MCP SERVER | NO CODE | CLOUD HOSTED

Shumei Anti-Fraud MCP for AI Agents

Automating Content Moderation and Fraud Detection in Real Time

Shumei Anti-Fraud brings professional risk assessment and anti-fraud capabilities directly into your AI agent. It analyzes text, images, audio clips, and device identifiers in real time. Use it to automatically filter out spam, detect malicious bots, and flag explicit or abusive content before it hits your platform.

A+ Quality Score 100/100

fraud-detection | nsfw-filtering | bot-detection | risk-control | content-moderation

MAINTAINED BY VINKIUS CLOUD FOR AI AGENTS · JULY 2026
vinkius.com/mcp/shumei-anti-fraud

SCAN TO VIEW

The connectivity layer between AI and the world's software.

Vinkius sits between AI and every application. All communication passes through Vinkius Cloud via the Model Context Protocol (MCP) — with governance, observability, and security at every layer.

Your AI Connections Run Through Vinkius Cloud

The world's largest managed MCP catalog

Vinkius is the connectivity layer where AI connects to the software your business already runs. We handle the hosting, the security, the credentials, the uptime — you get agents that actually do things.

We operate the world's largest managed MCP catalog. Major SaaS platforms, CRMs, databases, and cloud providers — running, monitored, production-ready. This MCP server is hosted and maintained by the Vinkius Cloud for AI Agents.

> *The agent doesn't manage credentials, doesn't manage uptime, doesn't manage security. Vinkius does.*
>
> — Architecture principle

Four Pillars of the Vinkius Runtime

01 — Security by design

Credentials stay encrypted at rest via AES-256. The AI agent never touches raw keys — they're injected into a sandboxed V8 isolate at runtime. Actions are logged, and connections have an emergency kill switch.

02 — Built on MCP Fusion

This MCP server was built with **MCP Fusion**, the open-source framework (Apache 2.0) that powers the entire Vinkius catalog. Schema-as-firewall strips undeclared fields, compiled PII redaction runs at zero overhead, and cryptographic lockfiles produce git-diffable audit trails.

03 — Deterministic observability

Eight immutable metrics per endpoint: request volume, p95 latency, error rate, active connections, cost attribution. A live payload feed logs every tool call with mutation detection.

04 — Autonomous operations

Servers are deployed, monitored, and patched autonomously. New capabilities and security patches ship weekly. Zero-downtime deployments ensure continuous availability across all managed MCP servers.

AES-256	Ed25519	24h TTL	V8 Isolate
Encryption at rest	PKI vault signatures	Ephemeral session keys	Sandboxed execution

One Token. Instant Access.

Every MCP server on Vinkius is accessed through a **Connection Token**. Tokens are generated in the cloud dashboard and produce a unique MCP endpoint URL. Paste this URL into any MCP-compatible client — no SDK required.

A single token can serve **multiple AI clients simultaneously**, or you can issue separate tokens per client for granular access control. Each token tracks its own request count, last activity timestamp, and can be individually enabled or revoked.

MCP ENDPOINT

```
https://edge.vinkius.com/{token}/mcp
```

Claude · Cursor · VS Code · Windsurf · Grok · Gemini

Security Is the Architecture

Security in Vinkius is not a feature — it's the foundation of the runtime. The gateway enforces multiple independent protection layers between AI agents and third-party APIs.

01 — Ed25519 PKI Vault

Every workspace has an Ed25519 Master Key. Session keys are generated ephemerally (24h TTL) and signed by the Master Key. Credentials never leave the vault boundary.

02 — V8 Isolate Sandboxing

Tool code runs inside isolated-vm V8 isolates with 64 MB memory caps and per-request timeouts. No filesystem access, no network access except through the SSRF-guarded fetch bridge.

03 — SSRF Guard

All outbound HTTP requests are DNS-resolved and validated before execution. Private IP ranges (10.x, 172.16-31.x, 192.168.x, AWS metadata 169.254.x) are blocked at the network layer.

04 — DLP & PII Redaction

A ResponseGuard pipeline intercepts every tool response. Configurable redaction patterns strip sensitive fields (emails, SSNs, card numbers) before data reaches the AI agent.

05 — Cryptographic Audit Trail

Every request is signed into a SHA-256 hash chain with Ed25519 signatures. Events form a tamper-proof, SIEM-exportable forensic record.

06 — Honeytoken Trap System

Phantom credentials are injected into isolated environments. If a honeytoken is used outside Vinkius infrastructure, the server is quarantined instantly.

Emergency Kill Switch

EU AI Act Art. 14(1) Compliant

The kill switch is an **emergency halt** mechanism — not a simple toggle. When triggered, it executes three actions atomically:

01 — Server deactivated

The MCP server is immediately taken offline across the entire cluster.

02 — All tokens revoked

Every connection token is invalidated. Total lockout — reconnection blocked until new tokens are issued.

03 — WebSocket connections killed

Active connections terminated via Redis pubsub broadcast. Propagates to every runtime node in the cluster.

Full Visibility. Zero Guesswork.

The Vinkius cloud dashboard includes a full MCP Governance suite — real-time analytics and security controls for production AI operations.

Control Plane

KPI dashboard with request volume, latency, success rate, token consumption, and AI-generated operational briefings.

Agent Activity

Which AI clients are connecting, how often, and what they're doing — real-time session tracking.

FinOps

Cost tracking per tool, payload compression savings, budget optimization signals, and consumption trends.

Tool Health

Slowest and most error-prone tools, with actionable root-cause insights and performance baselines.

Firewall & DLP

PII redaction activity, sensitive data protection counters, and security event timeline.

Incident Log

Error trends, failure rates, status-code breakdowns, and forensic audit trail access.

> Get started at **cloud.vinkius.com** — connect your AI agent in under 60 seconds.

Shumei Anti-Fraud MCP

4 tools available
Cloud-hosted on Vinkius

Shumei Anti-Fraud connects top-tier risk assessment tools to your workflow, letting your AI agent police itself from the inside. Instead of building complex internal filtering layers, you simply pipe suspicious data through this MCP. Your agent can automatically check incoming text for spam or abuse, scan uploaded avatars and images for restricted content, and even validate user IPs against known fraud databases. It's how you build safety features without becoming a full-time moderation company. Once connected via Vinkius, your agent gets access to this powerful layer of defense right alongside everything else it needs. You can handle bot detection—by checking device IDs or IP ranges—and moderate content across text, images, and audio streams all from one place.

Core Capabilities

01 — Check Text Risk

Scans blocks of submitted text to determine if they contain spam indicators, abuse, or explicit material.

02 — Analyze Image Content

Reviews uploaded images to detect restricted, NSFW, or policy-violating visual content.

03 — Predict Audio Risk

Assesses audio clips for potential risk flags, identifying prohibited material.

04 — Validate Device Identity

Determines if a provided device ID or IP address is flagged as fraudulent, potentially using a VPN or emulator.

One Click on Vinkius — From Prompt to Execution

Available at **vinkius.com/mcp/shumei-anti-fraud** — connect your AI agent in three steps.

01 Subscribe to this MCP and acquire your unique Access Key from the Shumei Control Panel.

02 Inject the provided key into your LLM logic as an automated moderation backend.

03 Your agent can now call specific tools, running real-time audits on user inputs like uploaded media or chat transcripts.

The bottom line is that you treat content safety checks just like any other function call in your existing AI workflow.

Built For

Community Managers and Trust & Safety Agents need this. If your platform relies on user-generated content (UGC) or needs to maintain a clean chat experience, you run into moderation nightmares. This MCP gives you the tools needed to stop bad actors before they mess up your community.

Community Manager

Automating the scanning of large chat transcripts for violations and spam flags across multiple users.

Trust & Safety Agent

Intercepting fraudulent account creations by checking IP risks or device IDs against known fraud databases.

Backend Developer

Prototyping safety wrappers around user-uploaded images, avatars, and other media types before they hit the main database.

What Changes When You Connect

01 You stop spam and abuse before they enter your platform. Use `check_text_risk` to vet all incoming messages, keeping your chat clean.

02 Block bad actors at the source. Running `check_device_risk` lets you validate IPs and device IDs, stopping VPN or bot activity instantly.

03 Review uploaded media safely. By using `check_image_risk`, you ensure that avatars and user-uploaded visuals meet policy standards.

04 Go beyond text. The ability to run `check_audio_risk` means your moderation coverage extends to voice content.

05 Build a robust safety layer quickly. This MCP lets developers wrap up complex security checks without building the underlying detection models themselves.

Real-World Applications

Cleaning up massive chat transcripts

A community manager needs to audit thousands of old messages for violations. They use their agent to run `check_text_risk` on every transcript block, instantly flagging all instances of spam or abuse for human review.

Preventing bot account creation

A Trust & Safety Agent detects a surge of new accounts. They use the agent to run `check_device_risk` on the IPs and device IDs associated with these accounts, filtering out all known emulator or VPN connections.

Vetting user profile pictures

A backend developer is building a new avatar upload system. Before saving any image, they run `check_image_risk` on the uploaded file, ensuring no NSFW content enters the database.

Moderating live audio streams

The platform allows users to submit voice messages for review. The agent intercepts these files and calls `check_audio_risk`, rejecting any clip that violates community standards.

Patterns to Avoid

Checking content manually

✗ AVOID

A team member copies a suspicious text block, pastes it into an external form, and waits for a manual review. This is slow and doesn't scale.

✓ INSTEAD

Integrate `check_text_risk` directly into your agent's flow. The AI calls the tool instantly on receipt of content, providing real-time flagging.

Ignoring device metadata

✗ AVOID

A developer assumes a new user is legitimate because they signed up via a simple form. They miss out on detecting bot farms using emulators.

✓ INSTEAD

Always run `check_device_risk` immediately upon sign-up. This verifies the source's legitimacy, blocking known fraudulent hardware.

Treating media checks as optional

✗ AVOID

The system allows images to be uploaded without checking them first. The platform ends up hosting inappropriate or restricted content.

✓ INSTEAD

Make `check_image_risk` a mandatory step in your upload pipeline. The agent fails the process if the image is flagged.

The Right Fit

Use Shumei Anti-Fraud MCP if your platform's primary concern is user-generated content safety, whether it's text, images, audio, or device identity. You need automated gatekeeping to handle spam, abuse, and policy violations at scale. Don't use this if you only deal with pre-vetted, internal data that never touches the public internet. If your goal is simple data storage without any moderation requirements, this MCP adds unnecessary complexity. However, if you suspect bot activity or need content screening across multiple media types (text via `check_text_risk`, images via `check_image_risk`, etc.), this is essential.

Shumei Anti-Fraud: Stopping Spam and Abuse in Community Chat Moderation

Today, moderation for chat transcripts means endless manual review. You have to copy suspect threads into separate tools, cross-reference user IDs, and wait hours or days for a human team to catch every instance of spam links or abusive language. This process is costly, slow, and always leaves gaps.

With this MCP, the whole process runs automatically inside your agent. When content hits your system, it instantly calls `check_text_risk`. Your agent gets back a clean flag—'Pass' or 'Reject/Spam'—allowing you to take action immediately without human intervention.

Shumei Anti-Fraud: Blocking Fraudulent Device Connections on User Signups

Without robust checks, your sign-up flow lets anyone through. Bots and bad actors can easily cycle through VPNs or use emulated devices to create multiple fake identities, skewing data and disrupting the community.

Now, integrate `check_device_risk` into your user authentication logic. Your agent verifies the source IP and device ID in milliseconds. If it fails the check, you block the signup immediately, protecting your integrity.

Shumei Anti-Fraud: 4 Tools for Content Moderation & Fraud Detection

Use these four tools to run deep risk audits on text, images, audio clips, and user connection metadata.

#	TOOL	DESCRIPTION
01	`check_audio_risk`	Predicts the level of risk associated with a given audio clip.
02	`check_device_risk`	Determines if a specific device ID or IP address is flagged as fraudulent, potentially using a VPN or emulator.
03	`check_image_risk`	Scans an image file to check for NSFW or restricted content types.
04	`check_text_risk`	Analyzes a piece of text to identify spam, abuse, or explicit material.

See It in Action

Real prompts you can use once this MCP is connected to your AI agent through Vinkius Cloud.

U Hey, scan this message for spam or abuse: 'Check out my crypto site! 💰 Link here...'?

Content Risk Report

- **Text Analysis:** Passed (Abuse), Failed (Spam)
- **Classification:** Advertising/Financial Scam
- **Action Recommended:** Block and warn user.

This content is highly likely to be spam due to the external link structure and promotional keywords.

U What's up with this device ID? It came from a new account.

Device Validation Status

- **Input ID:** `192.168.1.1` (Simulated)
- **Risk Score:** HIGH
- **Status:** Fraudulent/VPN Detected
- **Reasoning:** The IP range is associated with multiple known VPN exit nodes and has been flagged for emulator usage.

Do not allow this device to create an account.

U I need to check the image attached. Is it okay?

Processing image...

Shumei Image Audit Results:

- **NSFW Flag:** Negative (Clean)
- **Restricted Content:** None detected.
- **Confidence Score:** 98% Safe

The image is clean and passes all policy checks. You can let the user upload it.

Frequently Asked Questions

01 **How does Shumei Anti-Fraud help me detect bot accounts?**

It detects bots by checking device identifiers and IP addresses using `check_device_risk`. If the source is flagged as a VPN or emulator, you know it's suspicious and can block the account creation before it starts.

02 **Can I use this MCP to check user-uploaded photos for inappropriate content?**

Yes. Use `check_image_risk` to scan any uploaded image against a database of restricted or NSFW content, ensuring your platform remains safe and compliant.

03 **Is Shumei Anti-Fraud good for general chat moderation?**

It's excellent. You run `check_text_risk` on every message stream. This instantly flags spam, abuse, or explicit text so your agent can take action immediately.

04 **What if I need to moderate audio messages? Does Shumei Anti-Fraud handle that?**

It does. You use `check_audio_risk` to analyze the content of an audio clip, flagging it if it violates your community's standards before playback.

05 **Do I have to build custom models for fraud detection with Shumei Anti-Fraud?**

No. This MCP connects you to existing industry-leading risk databases. You just call the specific tool—like `check_device_risk`—and get a professional, real-time risk score back.

Go Live in 60 Seconds

Get your connection token from **cloud.vinkius.com**, then paste the endpoint URL into any MCP-compatible client.

YOUR MCP ENDPOINT

```
https://edge.vinkius.com/[TOKEN]/mcp
```

CLIENT	WHERE TO CONFIGURE
Claude AI	Profile → Customize → Connectors → "+" → Add custom connector → Paste endpoint
Cursor	Settings → Features → MCP Servers → "+ Add New MCP Server" → Type: SSE → Paste endpoint
VS Code	Ctrl/Cmd+Shift+P → "MCP: Add Server" → add `"shumei-anti-fraud": { "url": "…" }`
Windsurf	MCP Settings → `mcp_settings.json` → Add endpoint URL
ChatGPT	Settings → Tools & plugins → Add MCP server → Paste endpoint
Gemini	Extensions → Add MCP Server → Paste endpoint URL

ASK AN AI ABOUT THIS

Let your preferred AI explain this MCP server

- Ask ChatGPT
- Ask Claude
- Ask Perplexity
- Ask Gemini
- Ask Grok

READY TO CONNECT

Shumei Anti-Fraud is live on Vinkius Cloud.

Get your connection token, paste it into your AI agent, and start building. No SDK. No deployment. Just results.

Start at cloud.vinkius.com →

INDEPENDENT PLATFORM DISCLAIMER

Vinkius is an independent platform and is not affiliated with, endorsed by, sponsored by, verified by, or otherwise authorized by Shumei Anti-Fraud. All third-party trademarks, logos, and brand names are the property of their respective owners. Their use in this document is strictly for informational purposes to identify service compatibility and interoperability.

DOCUMENT INFORMATION

Generated	July 2026
MCP Server	Shumei Anti-Fraud MCP
Server ID	`019d8480-4d12-7158-a291-8a98af9bb9bb`
Platform	Vinkius Cloud for AI Agents
Endpoint	`https://edge.vinkius.com/{token}/mcp`

LICENSE & USAGE

This document is generated automatically by the Vinkius PDF Engine. Content reflects the MCP server configuration at the time of generation and may change as updates are deployed. For the most current information, visit vinkius.com/mcp/shumei-anti-fraud.